Microbial Diversity!

Ashley **Shade**

Michigan State University

28 June 2018

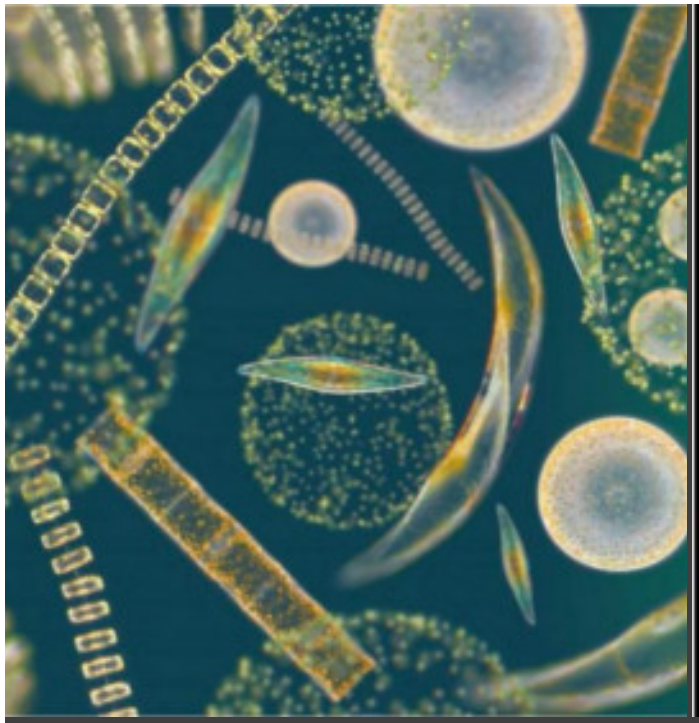
#EDAMAME2018

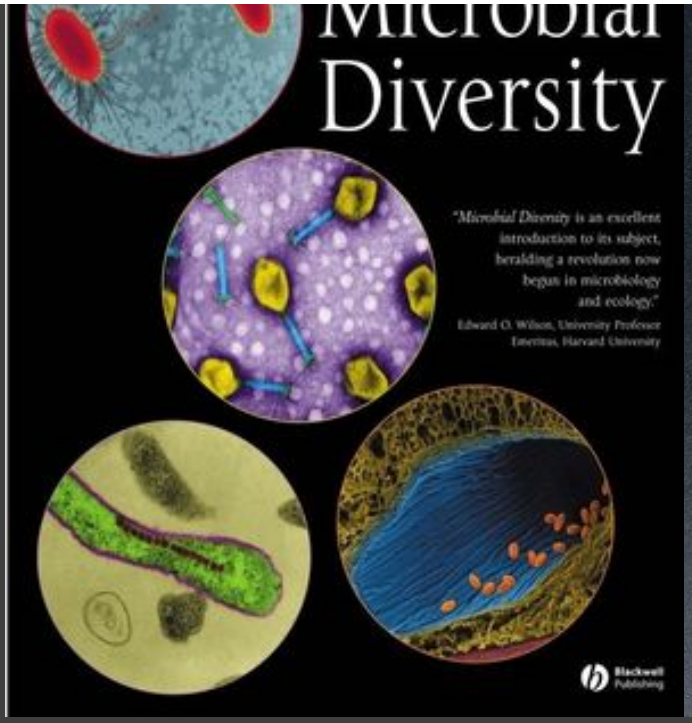
@ashley17061

shadeash@msu.edu

Outline

- What is diversity?
  - Alpha (within-sample) diversity
    - Challenges in quantifying alpha diversity & the macroecology perspective
  - Beta (comparative) diversity
    - Community resemblances
    - Ordinations
    - Hypothesis testing

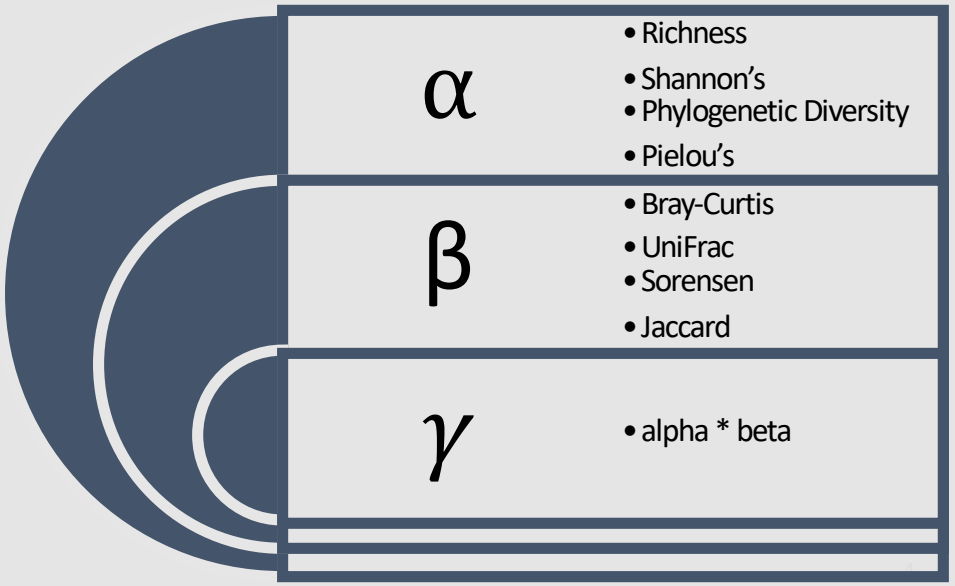




What is diversity?

Diversity in all of its glory

- **"Diversity" is a vague word.**
- In ecology, it has there are many types of diversity, and there are many components to that contribute to those types.



Whittaker introduces alpha, beta, gamma diversity (1972)

TAXON 21 (2/3): 213-251. MAY 1972

EVOLUTION AND MEASUREMENT OF SPECIES DIVERSITY*

*R. H. Whittaker***

Summary

Given a resource gradient (e.g. light intensity, prey size) in a community, species evolve to use different parts of this gradient; competition between them is thereby reduced. Species relationships in the community may be conceived in terms of a multidimensional coordinate system, the axes of which are the various resource gradients (and other aspects of species relationships to space, time, and one another in the community). This coordinate system defines a hyperspace, and the range of the space that a given species occupies is its niche hypervolume, as an abstract characterization of its intra-community position, or niche. Species evolve toward difference in niche, and consequently toward difference in location of their hypervolumes in the niche hyperspace. Through evolutionary time additional species can fit into the community in niche hypervolumes different from those of other species, and the niche hyperspace can become increasingly complex. Its complexity relates to the community's richness in species, its alpha diversity.

The confusion continues… for decades

A consistent terminology for quantifying species diversity?

Claudia E. Moreno · Pilar Rodríguez

Abstract There is a genuine need for consensus on a clear terminology in the study of species diversity given that the nature of the components of diversity is the subject of an ongoing debate and may be the key to understanding changes in ecosystem processes. A recent and thought-provoking paper (Jurasinski et al. Oecologia 159:15–26, 2009) draws attention to the lack of precision with which the terms alpha, beta, and gamma diversity are used and proposes three new terms in their place. While this valuable effort may improve our understanding of the different facets of species diversity, it still leaves us far from achieving a consistent terminology. As such, the conceptual contribution of these authors is limited and does little to elucidate the facets of species diversity. It is, however, a good starting point for an in-depth review of the available concepts and methods.

Keywords Alpha diversity · Beta diversity · Gamma diversity · Species richness · Turnover

Oecologia 2010

Ecology Letters 2011

Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist

Abstract

A recent increase in studies of β diversity has yielded a confusing array of concepts, measures and methods. Here, we provide a roadmap of the most widely used and ecologically relevant approaches for analysis through a series of mission statements. We distinguish two types of β diversity: directional turnover along a gradient vs. non-directional variation. Different measures emphasize different properties of ecological data. Such properties include the degree of emphasis on presence/absence vs. relative abundance information and the inclusion vs. exclusion of joint absences. Judicious use of multiple measures in concert can uncover the underlying nature of patterns in β diversity for a given dataset. A case study of Indonesian coral assemblages shows the utility of a multi-faceted approach. We advocate careful consideration of relevant questions, matched by appropriate analyses. The rigorous application of null models will also help to reveal potential processes driving observed patterns in β diversity.

Marti J. Anderson,[1]* Thomas O. Crist,[2] Jonathan M. Chase,[3] Mark Vellend,[4] Brian D. Inouye,[5] Amy L. Freestone,[6] Nathan J. Sanders, Howard V. Cornell,[8] Liza S. Comita,[9] Kendi F. Davies,[10] Susan P. Harrison,[8] Nathan J. B. Kraft,[11] James C. Stegen[12] and Nathan G. Swenson[13]

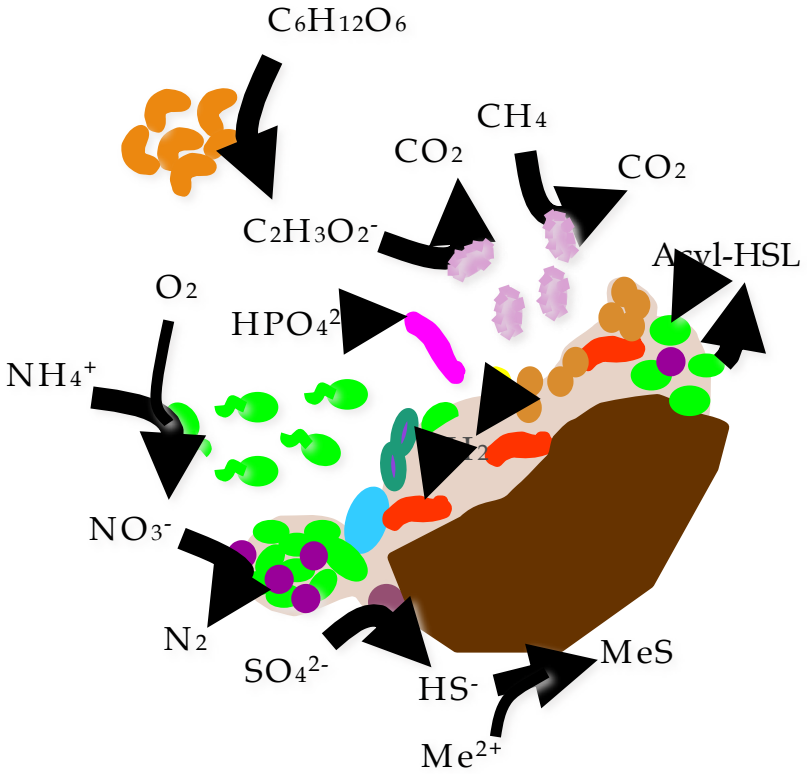
Within-sample *(aka alpha)* diversity

Richness: How many species* (operational taxonomic units)?

Evenness: How equivalent are species in their abundances?

Composition: Who is there/ what is the (relative) contribution of the species?

Phylogenetic diversity: What is the collective evolutionary breadth of species, inferred from phylogeny?



Combinations of the above components are used to calculate other diversities: Shannon diversity, Simpson, *etc.*

Also: extrapolations of alpha diversity, e.g. Chao

Alpha diversity

- There are many ways to calculate alpha diversity, and each emphasizes different aspects of highly dimensional community data
- Precision in the definition (and math!) used to calculate alpha diversity
- Not everyone needs to use the same alpha diversity metric
  - Scientifically justify decision
  - Consider complementary insights provided by multiple metrics
  - Make data available so that others can use their favorite metric

OPEN

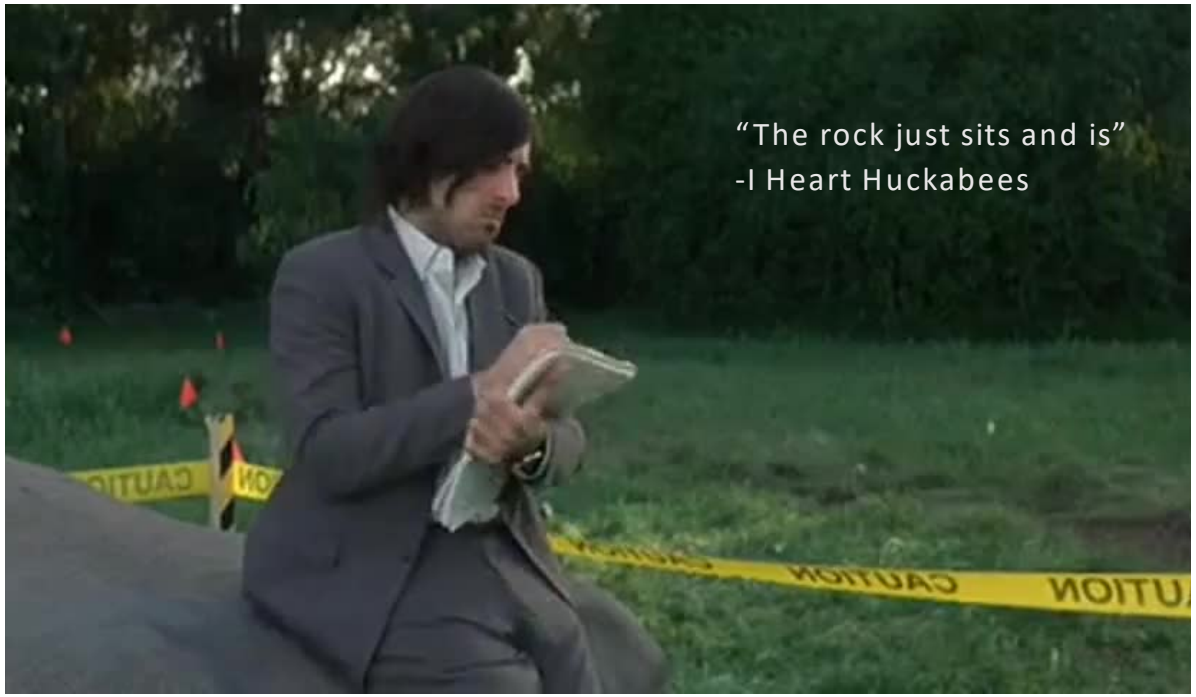
The ISME Journal (2017) 11, 1–6
© 2017 International Society for Microbial Ecology All rights reserved 1751-7362/17
www.nature.com/ismej

PERSPECTIVE

Diversity is the question, not the answer

Ashley Shade
Department of Microbiology and Molecular Genetics, Program in Ecology, Evolution, and Behavior, and DOE Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, MI, USA

Local diversity (within-sample or alpha diversity) is often implicated as a cause of success or failure of a microbial community. However, the relationships between diversity and emergent properties of a community, such as its stability, productivity or invasibility, are much more nuanced. I argue that diversity without context provides limited insights into the mechanisms underpinning community patterns. I provide examples from traditional and microbial ecology to discuss common complications and assumptions about within-sample diversity that may prevent us from digging deeper into the more specific mechanisms underpinning community outcomes. I suggest that measurement of diversity should serve as a starting point for further inquiry of ecological mechanisms rather than an 'answer' to community outcomes.
The ISME Journal (2017) **11,** 1–6; doi:10.1038/ismej.2016.118; published online 16 September 2016



- Diversity is the outcome of ecological processes, not a process in itself
- There is no "good" or "bad" diversity
- To think objectively about diversity, we need to let go of value judgement

Diversity has no inherent value


It is hard to withhold judgement

Challenges in quantifying and comparing diversity

1

There is no absolute diversity

2

Exhaustive observational effort is key but rarely achieved

3

Species or taxonomic units and their communities must be defined and ecologically meaningful

Challenges in quantifying and comparing diversity

1

There is no absolute diversity

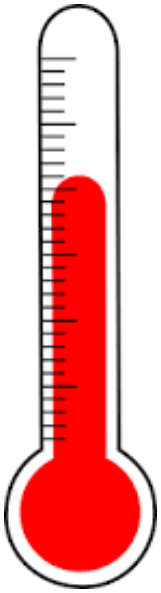
2

Exhaustive observational effort is key but rarely achieved

3

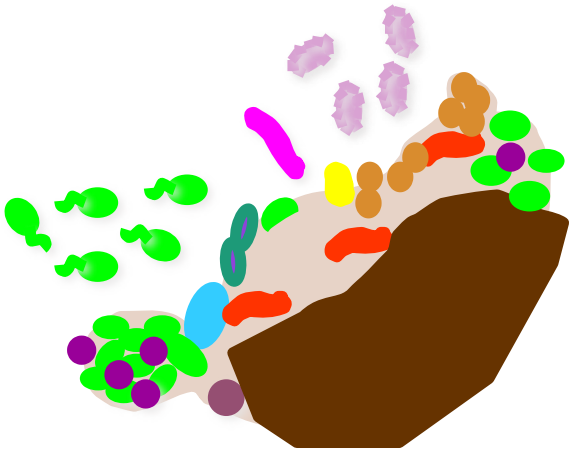
Species or taxonomic units and their communities must be defined and ecologically meaningful

We cannot know “absolute” diversity



100 ℃ = 212 F

1 cm = 0.394 in



Alpha diversity is in the eye of the beholder

Richness – which microbial ecologist is most accurate?

a) Sue makes microscope slides and counts cell morphotypes, which is then directly related to the total number of cells observed (community size)

b) Jane sequences the 16S rRNA gene and counts "OTUs" at 97% sequence identity and uses a reference database

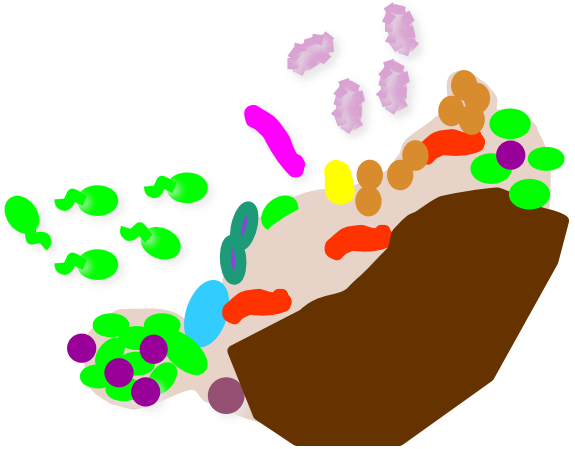
c) Sally sequences the 16S rRNA gene and counts "OTUs" at 97% sequence identity and clusters "de novo"

d) Danny sequences the 16S rRNA gene and then counts "exact sequence variants" (100% clustering)

e) Tim quantifies the single copy gene rpoB with qPCR and then sequences the amplicons to count

f) Tom only cares about the Streptomyces so he counts those only

g) Jesse uses five different FISH probes with microscopy



Alpha diversity is in the eye of the beholder

If observation of the community is not exhaustive, no one is accurate.

Challenges in quantifying and comparing diversity

1

There is no absolute diversity

2

Exhaustive observational effort is key but rarely achieved

3

Species or taxonomic units and their communities must be defined and ecologically meaningful

IDiv workshop on integrating microbes into macroecology, Leipzig Germany

JT Lennon, A Shade, K Küsel, P Keil, M Hermann, R Dunn, D Storch, B Bohannen, S Blowes, J Chase, N Sanders

What is macroecology?

- It is not the study of large charismatic macrofauna
- It is the study of large-scale patterns of diversity
  - **Scales across taxa : lineages, body sizes, energetic constraints**
  - **Scales over**
    to inter-cor
  - Scales ov

What are
diversity?
What are t
What are the

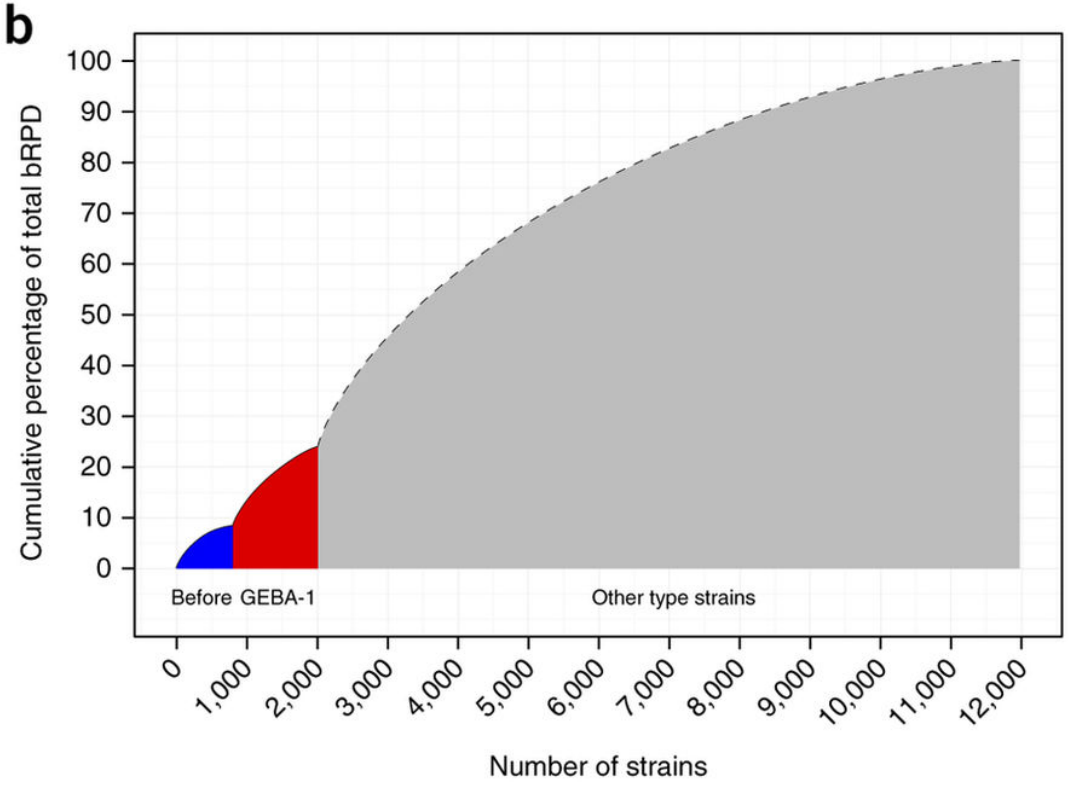
Observational effort

- How well have you sampled your community?
- How representative is your sampling effort of the true community?

How representative?
How do we observe communities of birds ?
(or other charismatic macrofauna)?
Is that method representative of the “true” community?

Under-sampling and unknowns:
There are no representative genomes for the vast majority of microbial diversity

(Balanced relative phylogenetic diversity)



- Reference-based

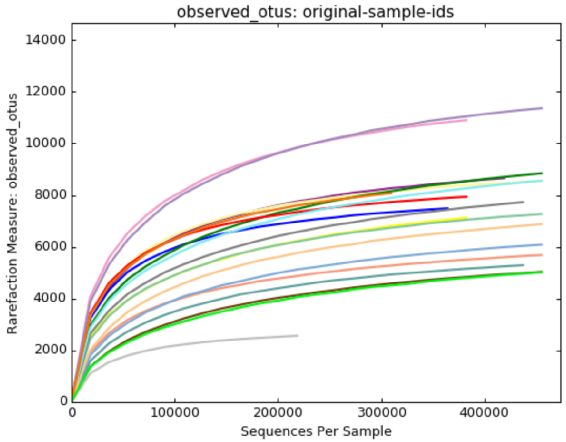
(e.g., using a curated database of sequence) species definitions are a gross underestimate

Mukherjee et al. 2017 Nature Biotechnology

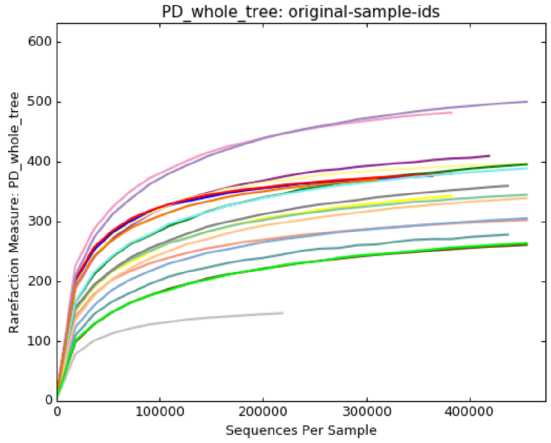
Subsampling sequencing datasets: Get "even"

- Because of sequencing artifacts and experimental design, there can be quite a range of quality sequences returned for each sample.
- Sub-sampling of sequences to achieve an even number across all samples within a dataset allows for comparing diversity across samples.
- **This is very important for being able to compare diversity across samples.**
  - Analogy: You are a tree ecologist. Would it be reasonable to directly compare the forest diversity (assessed by counting different types of trees) in a 1x1 m plot to a 1000 x 1000 km plot? The second plot has 1000x the coverage as the first and thus the comparison is unsound.
  - This is a question of the observational effort. We can only compare communities observed at an equal observational effort.
- Choose a sequencing depth that maximizes the number of samples that can be included at the most informative sequencing depth.
- If you have obvious outliers in sequencing depth (e.g., the median depth is 70K and you have one sample with 1000 sequences), get rid of it and save yourself heartache.

A



B



Rarefaction for estimating diversity given coverage

- I call this “exploratory” rarefaction
- It lets us understand how well we’ve observed our community
- Calculate a diversity of choice (e.g., richness)
- The question asked: Given an increase in sequencing (observation/coverage) effort, how likely are we to observe a new taxon?
- Exploratory rarefaction happens to understand your coverage BEFORE subsampling to an even sequencing effort

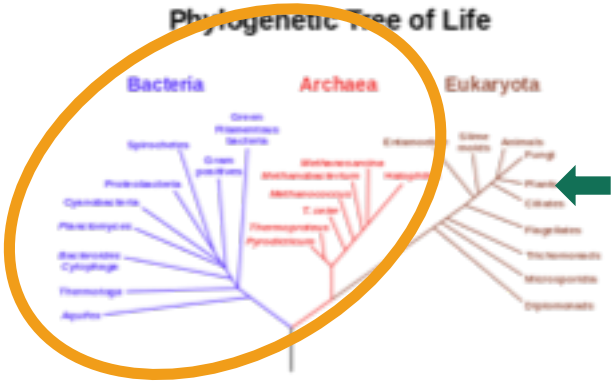
WARNING:

- There is no normalization method can make up for a bad run or low quality data!

Microbial ecologists sample communities destructively

- There are limitations in being able to go back to double check or re-assess as methods improve
- For sequencing methods, the nucleic acid extraction protocol is the first large bias
- The pooling of sample for microbial community analysis has been equated to putting the rainforest in a blender

What is a community?



2 of life's "domains" from a core of soil?

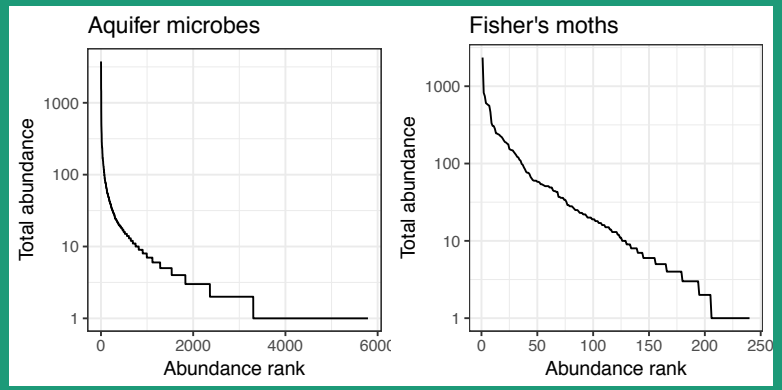
Trees on an island?

Microbial ecologists tend to throw out rare taxa, especially singletons

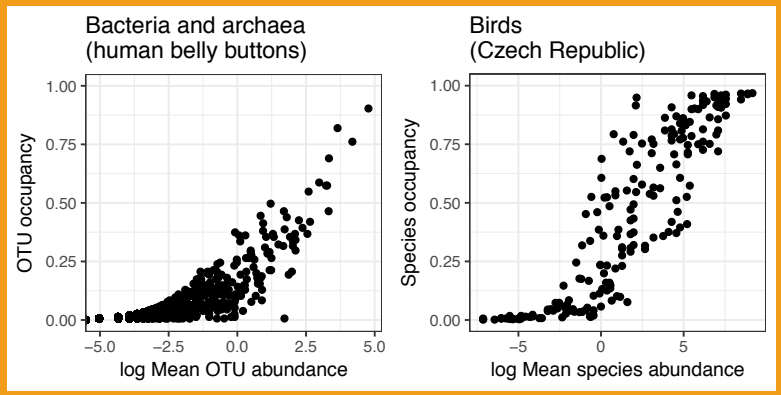
- A “complete” assessment, including rare taxa and especially singletons, are needed to fit models to understand underlying processes – a key goal in ecology

Rare taxa are important for understanding macroecological patterns and processes

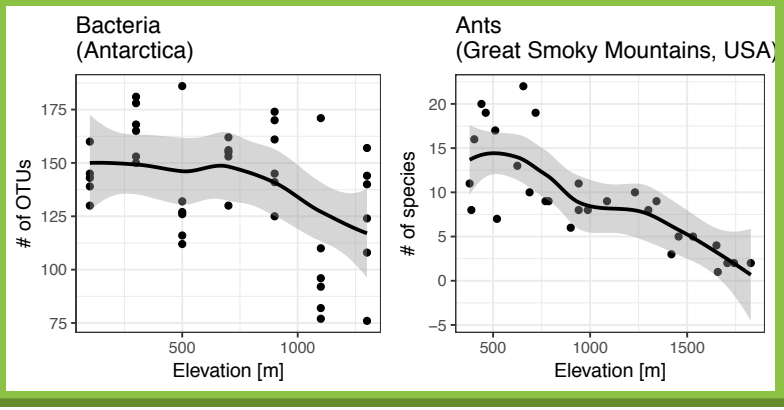
Species abundance distributions



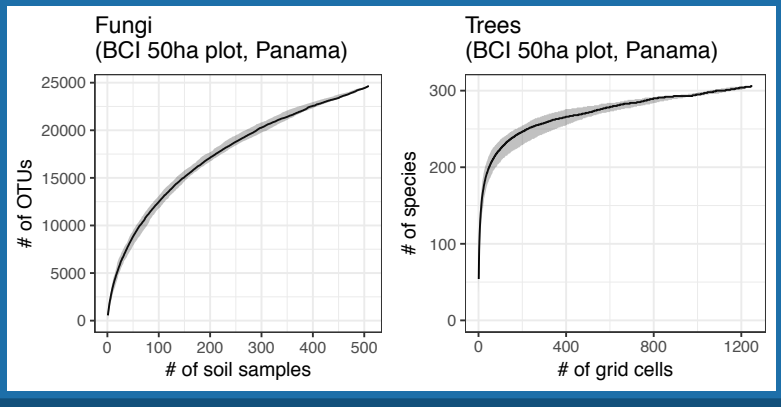
Abundance-occupancy (aka neutral models)



Diversity changes over gradients



Species area relationships



Microbial ecologists tend to throw out rare taxa, especially singletons

Motivation:

- Sequencing errors that are unknown/unquantified
- Superstition: we just don't trust the data

Scraping the bottom of the barrel: are rare high throughput sequences artifacts?

Shawn P. BROWN[a,*], *Allison M. VEACH*[a], *Anne R. RIGDON-HUSS*[b], *Kirsten GROND*[a], *Spencer K. LICKTEIG*[a], *Kale LOTHAMER*[a], *Alena K. OLIVER*[a], *Ari JUMPPONEN*[a]

[a]*Division of Biology, Kansas State University, 116 Ackert Hall, Manhattan, KS 66506, USA*
[b]*Department of Grain Science and Industry, Kansas State University, 201 Shellenburger Hall, Manhattan, KS 66506, USA*

ARTICLE INFO

Article history:
Received 25 March 2014
Revision received 26 June 2014
Accepted 19 July 2014
Available online 5 October 2014
Corresponding editor:
Marie Louise Davey

Keywords:
Fungi
High-throughput sequencing
Rare biosphere
Singleton

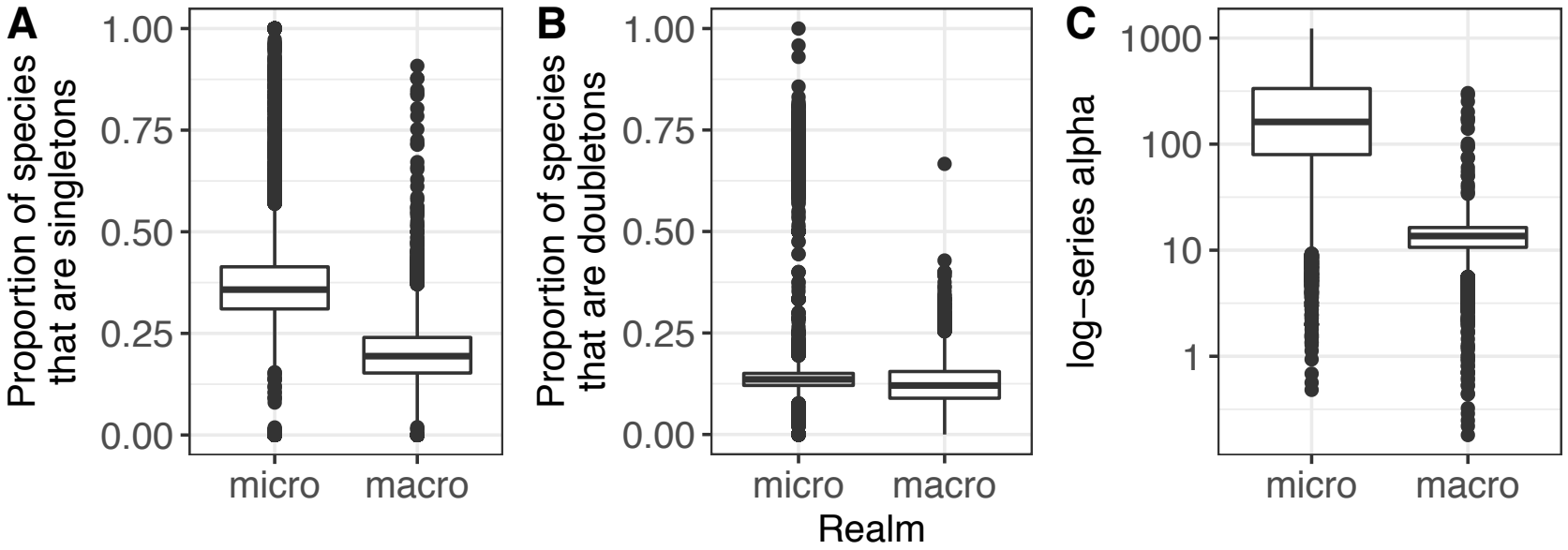
ABSTRACT

Metabarcoding data generated using next-generation sequencing (NGS) technologies are overwhelmed with rare taxa and skewed in Operational Taxonomic Unit (OTU) frequencies comprised of few dominant taxa. Low frequency OTUs comprise a rare biosphere of singleton and doubleton OTUs, which may include many artifacts. We present an in-depth analysis of global singletons across sixteen NGS libraries representing different ribosomal RNA gene regions, NGS technologies and chemistries. Our data indicate that many singletons (average of 38 % across gene regions) are likely artifacts or potential artifacts, but a large fraction can be assigned to lower taxonomic levels with very high bootstrap support (~32 % of sequences to genus with ≥90 % bootstrap cutoff). Further, many singletons clustered into rare OTUs from other datasets highlighting their overlap across datasets or the poor performance of clustering algorithms. These data emphasize a need for caution when discarding rare sequence data *en masse*: such practices may result in throwing the baby out with the bathwater, and underestimating the biodiversity. Yet, the rare sequences are unlikely to greatly affect ecological metrics. As a result, it may be prudent to err on the side of caution and omit rare OTUs prior to downstream analyses.

© 2014 Elsevier Ltd and The British Mycological Society. All rights reserved.

Fungal Ecology, 2015

Singletons are expected in ecological communities



Shade et al. in prep.

Institute for Biodiversity Working Group on Macro-Micro Ecology

Embrace those singletons 

Sometimes, to address specific questions, rare and singleton taxa are not needed…

- For example, comparative (beta) diversity using a weighted analysis – the singleton taxa are inconsequential

…but if your question is about alpha diversity, then **all high-quality, non-suspect data, even singletons, must be retained**

Our interpretation of discipline-biased observational effort issues:

Ecologists studying small organisms are too paranoid

Ecologists studying large organisms are too smug

Challenges in quantifying and comparing diversity

1

There is no absolute value of diversity

2

Exhaustive observational effort is key but rarely achieved

3

Species or taxonomic units and their communities must be defined and ecologically meaningful

What is a microbial species?

- Very likely, not an “OTU”

Past and future species definitions for *Bacteria* and *Archaea*

Ramon Rosselló-Móra[a,*], Rudolf Amann[b]

[a] Marine Microbiology Group, Department of Ecology and Marine Resources, Mediterranean Institute for Advanced Studies (IMEDEA, CSIC-UIB), E-07190 Esporles, Illes Balears, Spain
[b] Max Planck Institute for Marine Microbiology, D-28359 Bremen, Germany

ARTICLE INFO

Keywords:
Species concept
Species definition
Genomics

ABSTRACT

Species is the basic unit of biological diversity. However, among the different microbiological disciplines there is an important degree of disagreement as to what this unit may be. In this minireview, we argue that the main point of disagreement is the definition (i.e. the way species are circumscribed by means of observable characters) rather than the concept (i.e. the idea of what a species may be as a unit of biodiversity, the meaning of the patterns of recurrence observed in nature, and the why of their existence). Taxonomists have defined species by means of genetic and expressed characters that ensure the members of the unit are monophyletic, and exhibit a large degree of genomic and phenotypic coherence. The new technologies allowing high-throughput data acquisition are expected to improve future classifications significantly and will lead to database-based taxonomy centered on portable and interactive data. Future species descriptions of *Bacteria* and *Archaea* should include a high quality genome sequence of at least the type strain as an obligatory requirement, just as today an almost full-length 16S rRNA gene sequence must be provided. Serious efforts are needed in order to re-evaluate the major guidelines for standard descriptions.

© 2015 Elsevier GmbH. All rights reserved.

Systematic and Applied Microbiology
2015

This article gives a great historical background to microbial species definitions!

A final word about alpha diversity

- Higher diversity isn't necessarily better or healthier.
- Diversity is the outcome of ecological processes, not an ecological process in itself.

Comparative (beta) diversity

- Community resemblance
- Ordinations
- Hypothesis testing
- Environmental Gradients

When to use Comparative (beta) diversity

- Space / Time
- Categories (e.g., fire-affected, recovered)
- Gradients/empirical measurements (e.g., pH, temperature, chemistry)

- Look forward to Stuart's R lecture on category/gradient analyses!

Comparative diversity requires a measure of pair-wise community **resemblance**

- Resemblance = distance, similarity, dissimilarity
- Important decisions in choosing a resemblance metric:
  - Weighted v. Unweighted
  - Phylogenetic v. Taxonomic
- All pairs of resemblances are included in a sample by sample **resemblance (distance/similarity) matrix**
  - Simplifies the data and the analysis
- Choice of resemblance metric **will influence the outcome of community analysis**

Calculating resemblance: Bray-Curtis Example

$$d_{jk} = (sum\ abs(x_{ij}-x_{ik})/(sum\ (x_{ij}+x_{ik}))$$

Where d_{jk} is the Bray-Curtis index between samples j and k and x is the (relative) abundance of taxa I

See Legendre and Legendre book: *Numerical Ecology.* Chapter 7: "Ecological resemblance" for a comprehensive discussion of All the Resemblances Ever.

Making a Resemblance Matrix

1. OTU table (usually relativized)

	Soil 1	Soil 2	Soil 3
OTU 1	0	0.966	0.179
OTU 3	0.047	0.002	0.039
OTU 3	0.953	0.032	0.782

2. Chose appropriate resemblance (*e.g.,*

Bray Curtis, UniFrac)

3. Create a square (observation x observation)

resemblance matrix from pair-wise comparisons.

	Soil 1	Soil 2	Soil 3
Soil 1	0		
Soil 2	0.966	0	
Soil 3	0.179	0.787	0

Examples of Resemblance metrics

Unweighted metrics Weighted metrics

Metric name / *Accounts for*	Sørensen	Bray-Curtis	Weighted UniFrac	Unweighted UniFrac
Composition	X	X	X	X
OTU abundances?		X	X	
Phylogenetic diversity?			X	X

We can compare different distance/similarity measures to deduce the most important components of community structure for the overarching patterns observed

(This does not mean we “pick” the best ordination for our hypothesis - *confirmation bias*)

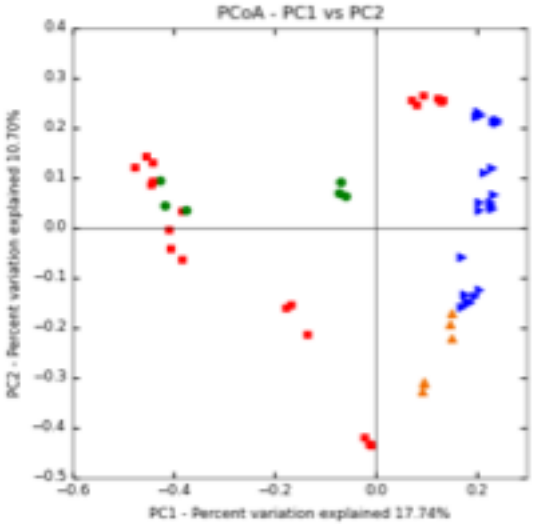
Weighted

Unweighted

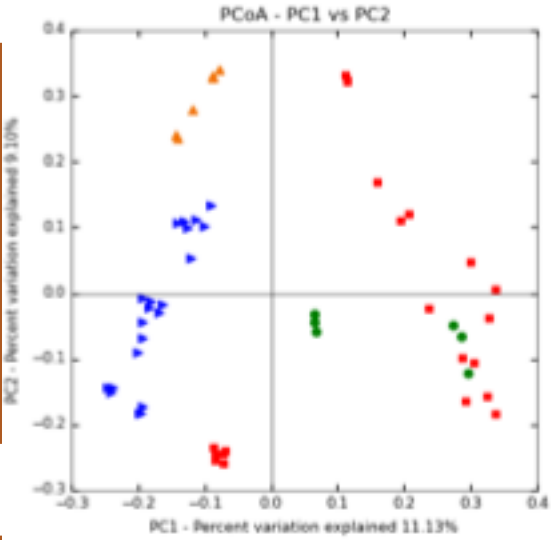
Taxonomic

Phylogenetic

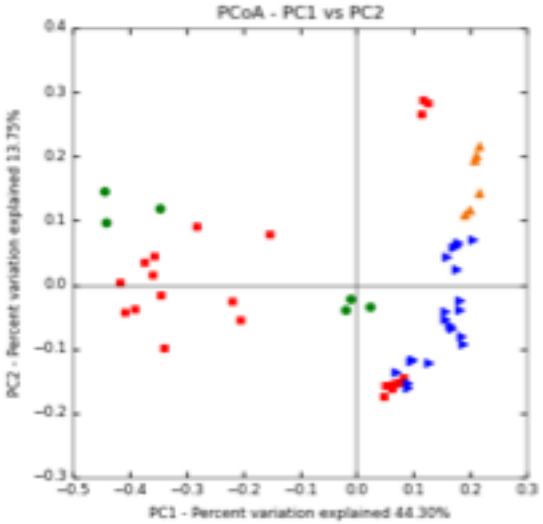
Bray-Curtis: 17.74% Ax 1



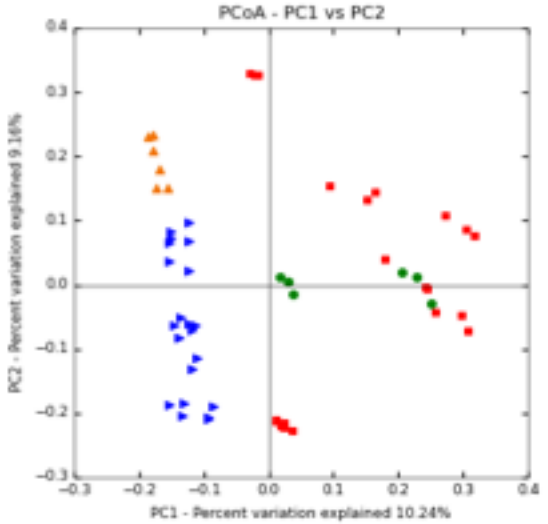
Sorenson: 11.13% Ax 1



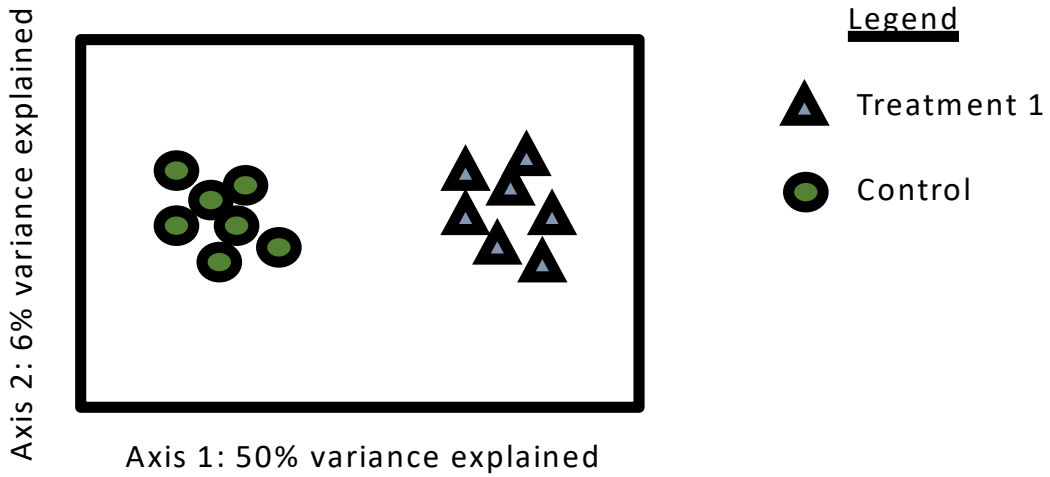
Weighted UniFrac: 44.3% Ax1



Unweighted UniFrac: 10.24% Ax1



Visualizing communities: Ordination



2 or 3 dimensional representation of the data

Each symbol is one community (compared by the chosen resemblance metric)

The distance between symbols represents the extent of differences between communities

First axis often explains most variance in the data, variation explained should be provided.

Types of ordinations

- **Non-metric multidimensional scaling (NMDS)**
  - Makes fewest assumptions, rank based, non-metric so cannot subsequently correlate to environmental measures

- **Principal coordinates analysis (PCoA)**
  - Assumes linear response of taxa, can use any resemblance
  - Special case Principal Components Analysis (PCA): based on Euclidean distance and not appropriate for communities (for environmental variables, PCA is okay)
  - Constrained version: Redundancy Analysis (RDA): variation in communities are *constrained* to be explained only by the measured variables. This is only okay if the ordination of the PCoA matches the RDA, which means that the important/most explanatory environmental variables were measured. Constrained analyses can be used for variance partitioning.

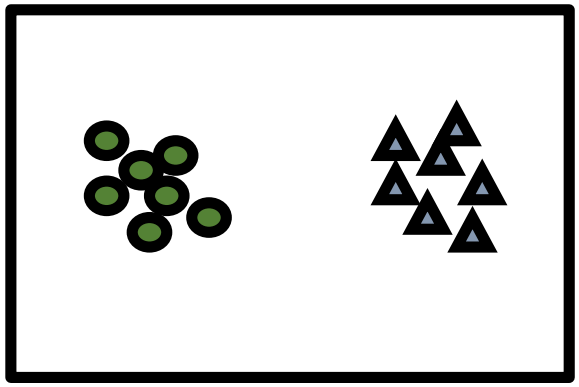
- **Correspondence analysis (CA)**
  - Assumes unimodal response of taxa, uses chi-squared distance
  - Constrained version: Constrained (Canonical) Correspondence Analysis (CCA)
  - Detrended correspondence analysis- if there is a large "horseshoe" or "arch" affect because too much variability is being squished onto too few dimensions; not often advisable.

- Avoid: Principle components analysis (PCA) for communities, Redundancy analysis (RDA) in situations that do not have a strong gradient (e.g., time) and detrended analyses *unless you really know what you are doing.*

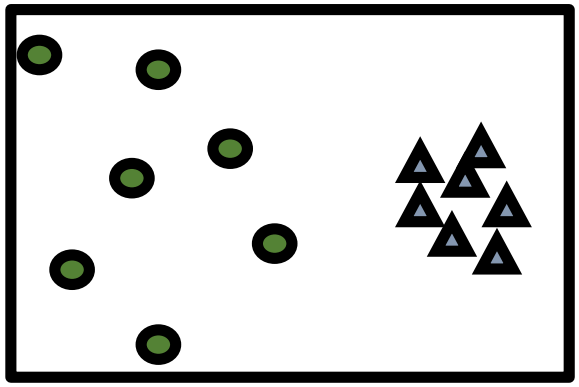
How do we interpret ordinations?

Think about: **CENTROID (mean)** or **DISPERSION (spread, variability)**

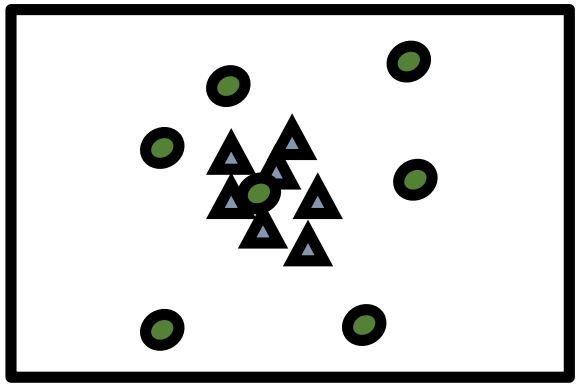
A. Different centroid, same spread



B. Different centroid, different spread



C. Same centroid, different spread



Describing patterns: Clusters & Gradients

Clusters = Are groups different? (*e.g.*, Treatment v. Control)

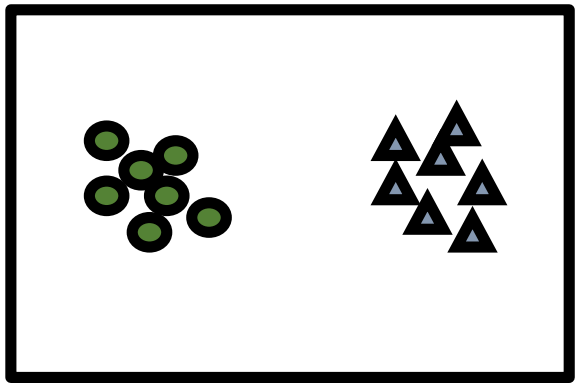
Also called: factors, qualitative variables

Gradients = Do communities change with known environmental changes? (*e.g.,* over time?)
Also called: continuous, quantitative, vector variables

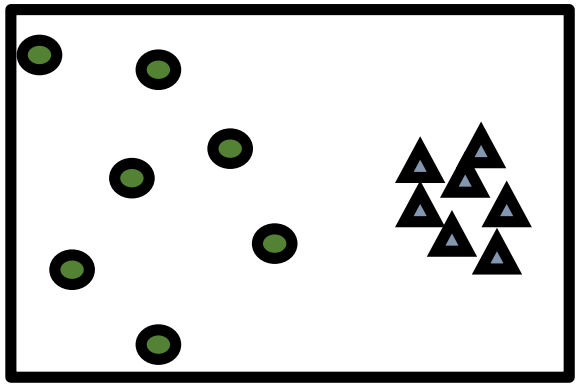
How do we interpret ordinations?

Think about: **CENTROID (mean)** or **DISPERSION (spread, variability)**

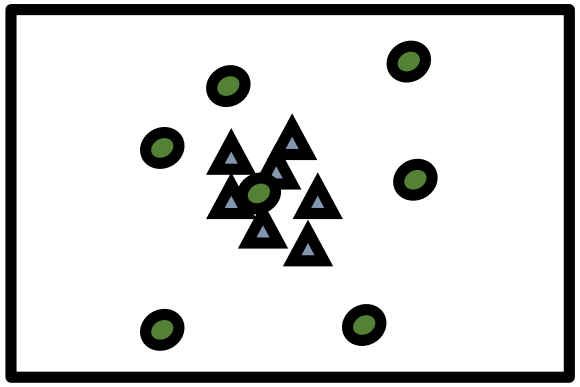
A. Different centroid, same spread



B. Different centroid, different spread



C. Same centroid, different spread

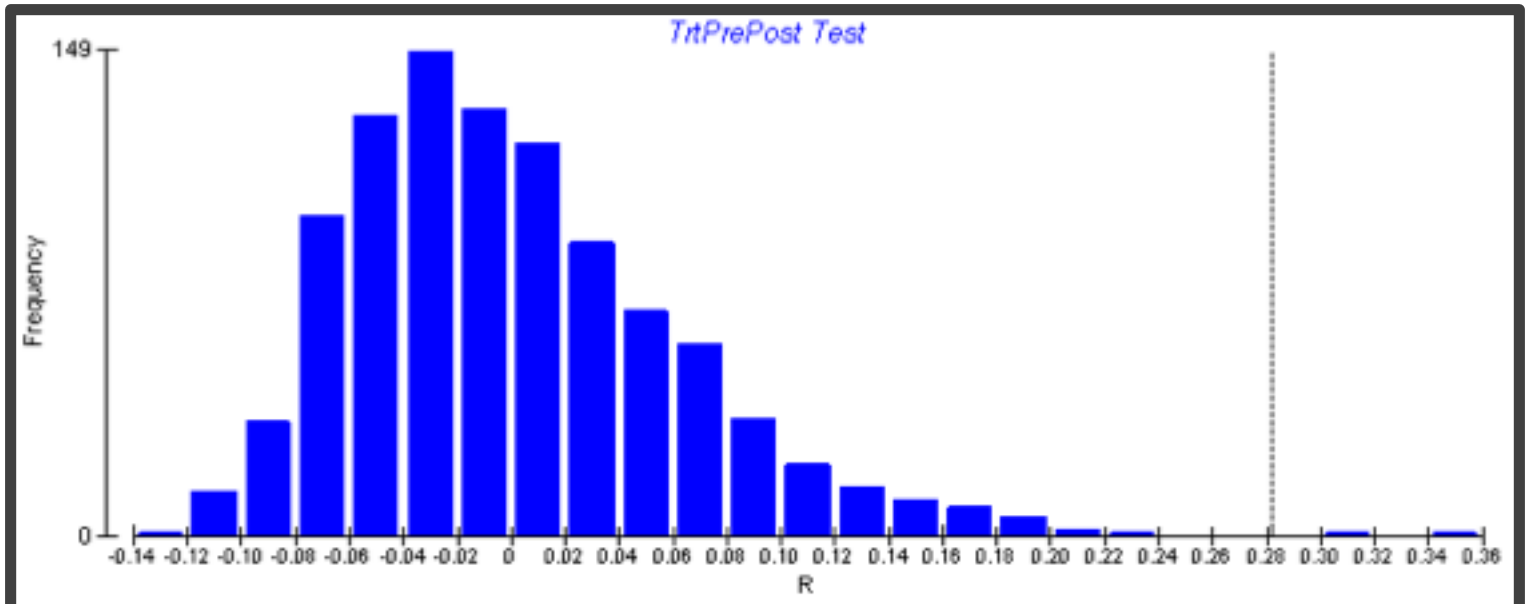


Non-parametric hypothesis tests

Non-parametric tests are used to test hypotheses of multivariate data when the underlying distribution of the data is unknown.

Non-parametric tests randomly re-sample the dataset to create a re-shuffled distribution, calculate a test statistic for each random distribution, and then ask the probably of finding the *actual* statistic given the random re-sampling distribution of the data.

It is important to use these tests for microbial beta diversity, as the assumptions of underlying normal distributions of most parametric tests (e.g., ANOVA) are violated.



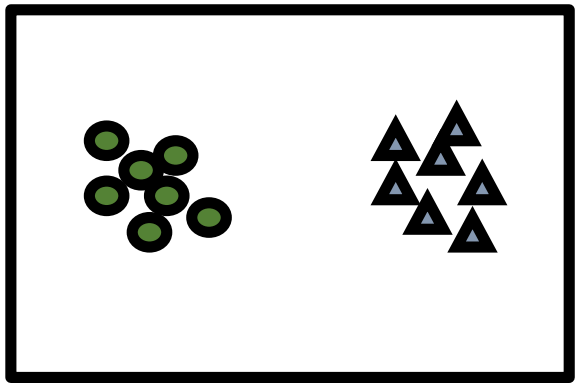
Hypothesis tests

- **Analysis of similarity (ANOSIM)**: rank based, least sensitive, makes fewest assumptions, permuted p-value
- **Multi-Response Permutation Procedure (MRPP)**: metric, permuted p-value
- **Permuted analysis of variance (PERMANOVA)**: assumes pseudo-F distribution, can accommodate a range of ANOVA-type experimental designs
- **Permutated analysis of dispersion** (**PERMDISP)**: p-values from permutation of residuals – tests specifically for differences in dispersion around centroid

Clusters: Testing for differences in *a priori* groups

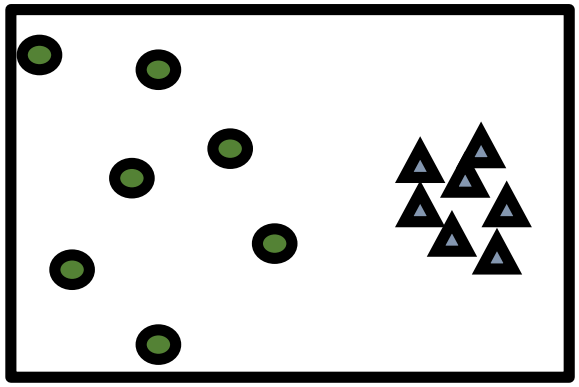
Permutation-based analyses to test hypotheses about group differences in **CENTROID (mean)** or **DISPERSION (spread, variability)**

A. Different centroid, same spread

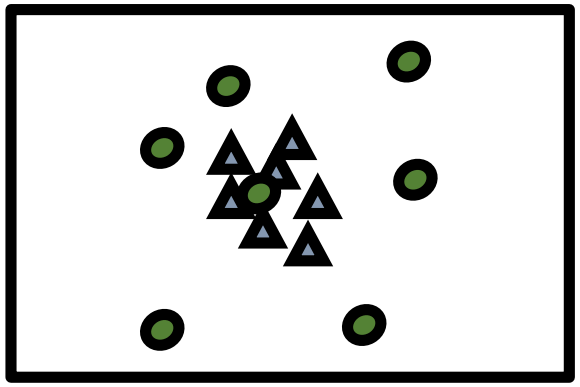


Test name	Centroid (mean)	Spread (variability)
PERMANOVA	X	X
MRPP	X	X
ANOSIM	X	X
PERMDISP		X

B. Different centroid, different spread



C. Same centroid, different spread



Team Ordination Activity